RECURSIVE ANALYSIS OF LINKED DATA FILES

W. Winkler, Bureau of the Census, and F. Scheuren, George Washington University

ABSTRACT

This paper demonstrates a methodology for analyzing two or more files when the only common information is name and address that is subject to significant error. Such a situation might arise with lists of businesses. We assume that a small proportion of records can be accurately matched. With the matched pairs we build an edit/imputation model and add predicted quantitative values, via a regression analysis to each file. Matching is then repeated with the common quantitative data and with name and address information. If necessary, the edit/impute, regression, and matching steps can be repeated in a recursive fashion. In large measure the ideas of Neter, Maynes, and Ramanathan (1965) are revised but with new tools.

KEYWORDS

Edit, Imputation, Record Linkage, Regression Analysis, Recursive Processes

1. INTRODUCTION

To make the best decisions, researchers and policymakers often need more information than is available in a single data base or in summary statistics from multiple files. Sound joint analyses of microdata records from two or more data files obviously require accurately linking records from one file to the corresponding records in the other file. This is possible if unique, common identifiers are available (e.g., Oh and Scheuren 1975; Jabine and Scheuren 1986). Probabilistic matching techniques also have been successful in many cases when sufficient overlapping information existed, even if subject to some error (e.g., Newcombe et al. 1959, Fellegi and Sunter 1969; Tepping 1969). However, until recently (e.g., Winkler and Scheuren 1995), the use of just highly error-prone, nonunique name and address information has not been sufficient for accurately analyzing matched microdata records.

> ***An example where microdata from multiple files might be needed is the following. One government agency maintains a file of the raw products used as inputs in a given industry while another agency independently maintains a file of the quantities of goods produced by those industries. For policy purposes, economists might wish to analyze inputs and outputs with company-specific microdata.***

The chief tools making the recent progress possible are advances in edit/imputation (**EI**) and record linkage (**RL**):

- Edit/imputation (**EI**) has traditionally been used to clean up data files in a manner that produces better summary statistics and, sometimes, allows more accurate microdata analyses. While some problems exist, knowledgeable analysts are the best starting point for defining the ways to edit (*i.e.*, clean-up) and impute (*i.e.*, revise) existing microdata. New edit software based on the model of Fellegi and Holt (1976) has fundamental advantages because it (1) checks the logical consistency of the entire edit system prior to the receipt of data, (2) in one pass, determines the optimal set of data to change so that a record satisfies all edits, and (3) consists of reusable software routines with the control of edit-based maintainable tables. Fellegi-Holt systems replace **ad hoc**, data-base-specific **if-then-else** routines that must be written from scratch in each application. Further, Fellegi-Holt edits are much more easily tied in with valid statistical imputation procedures and have the potential to allow important new statistically valid analyses of the microdata.

- Record Linkage (**RL**) has traditionally been used for linking files via difficult-to-use name and address information. **RL** was originally introduced by Newcombe (Newcombe *et al.* 1959) and mathematically formalized by Fellegi and Sunter (1969). In earlier work (Scheuren and Winkler 1993), we demonstrated an **RL** adjustment procedure that reduced bias due to record linkage error in regression analyses. Other recent work (Kim and Winkler 1995) has applied powerful new methods for linking files using quantitative data with or without corresponding name and address information.

In our present application, a four-step recursive approach is employed that is straightforward to carry out and very powerful. To start the process, we use an enhanced **RL** approach (*e.g.*, Winkler 1995, Belin and Rubin 1995) to delineate a set of pairs of records in which the matching error rate is estimated to be very low. A regression analysis, **RA,** is attempted. Then, we use an **EI** model developed on the low-error-rate linked records to edit and impute outliers in the remainder of the linked pairs. Another regression analysis (**RA**) is done and this time the results are then fed back into the linkage step so that the **RL** step can be improved (and so on). The cycle continues until the analytic results desired cease to change. Schematically, we have

$$\begin{tikzcd} & \mathrm{RA} \arrow[dr] & \\ \mathrm{RL} \arrow[ur] & \mathrm{RA} \arrow[l] & \mathrm{EI} \arrow[l] \end{tikzcd}$$

While each of these technologies is already well known, the edit portion of **EI** and much of **RL** are not generally used by statisticians. In this paper, we provide a means of integrating the methods to increase their utility even further. The integration presently depends primarily on statistical methods and not deep knowledge of computer science or operations research.

Organizationally, the material is divided into five sections, beginning with this brief introduction. In the second section, we give a little background on edit/imputation and record linkage technologies. The empirical data files constructed and the regression analyses undertaken are described in Section 3. In the fourth section, we present results. The final section consists of some conclusions and areas for future study.

2. EI AND RL METHODS REVIEWED

In this section, we undertake a short review of Edit/Imputation (**EI**) and Record Linkage (**RL**) methods. Our purpose is not to describe them in detail but simply to set the stage for the present application. Because Regression Analysis (**RA**) is so well known, our treatment of it is covered only in the particular application (Section 3).

2.1. Edit/Imputation

Methods of **editing** microdata have traditionally dealt with logical inconsistencies in data bases (*e.g.,* Nordbotten 1963, Granquist 1984). Early software consisted of **if-then-else** rules that were database-specific and very difficult to maintain or modify. Imputation methods were part of the set of **if-then-else** rules and could yield revised records that still failed edits (Sande 1982). In a major theoretical advance that broke with prior statistical methods, Fellegi and Holt (1976) introduced operations-research-based methods that both provided a means of checking the logical consistency of an edit system and assured that an edit-failing record could always be updated with imputed values so that the revised record satisfies all edits. An additional advantage of Fellegi-Holt systems is that their edit methods tie directly with current methods of **imputing** microdata (*e.g.,* Little and Rubin 1987).

Although we will only consider continuous data in this paper, **EI** techniques also hold for discrete data and combinations of discrete and continuous data. In any event, suppose we have continuous data. In this case a collection of edits might consist of rules for each record of the form

$$c_1X \prec Y \succ c_2X$$

In words,

If Y less than c_1X and greater than c_2X, then the data record should be reviewed.

Here Y may be total wages, X the number of employees, and c_1 and c_2 constants such that $c_1 < c_2$.

While Fellegi-Holt systems have theoretical advantages, implementation has been very slow because of the difficulty in developing general set covering routines needed for implicit-edit generation and integer programming routines for error localization (*i.e.*, determining the minimum number of fields to impute).

The current general Fellegi-Holt systems that run on a variety of computers consist of Statistics Canada's GEIS (Generalized Edit and Imputation System) for linear inequality edits (*e.g.*, Kovar, Whitridge, and MacMillan 1991), the Census Bureau's new SPEER (Structured Programs for Economic Editing and Referral) system for ratio edits of continuous data (*e.g.,* Winkler and Draper 1996), and the Census Bureau's DISCRETE system for edits of general discrete data (*e.g.*, Winkler and Petkunas 1996). Major improvements in Chernikova's algorithm were needed to make the GEIS system fast enough for real time use (Chernikova 1964 and 1965; Rubin 1975;

Filion and Schiopu-Kratina 1993). DISCRETE uses still other advances in operations research and computer science (Winkler 1995) that greatly reduce the redundant computation of general integer programming.

2.2. Record Linkage

A record linkage process attempts to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.*, 1959; Newcombe *et al* 1992), Fellegi and Sunter (1969) considered ratios **R** of probabilities of the form

$$\mathbf{R} = \mathbf{Pr}\ (\gamma \in \Gamma \mid \mathbf{M}) / \mathbf{Pr}\ (\gamma \in \Gamma \mid \mathbf{U})$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur. The fields compared (surname, first name, age) are called matching variables.

The decision rule is given by

If R > *Upper*, then designate pair as a link.

If *Lower* ≤ R ≤ *Upper*, then designate pair as a possible link and hold for clerical review.

If R < *Lower*, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that this decision rule is optimal in the sense that for any pair of fixed bounds on **R**, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds, *Upper* and *Lower*, are determined by the error bounds. We call the ratio **R** or any monotonely increasing transformation of it (typically a logarithm) a matching weight or total agreement weight.

Like **EI** methods, **RL** techniques have made major advances primarily as a result of new and much faster algorithms for comparing and utilizing information. Cheap, available computing and portable software has also had a role. Over about the last decade, there has been an outpouring of new work on record linkage techniques (*e.g.*, Jaro 1989, Newcombe, Fair, Lalonde 1992, Winkler 1995). Some of these results were spurred on by a series of conferences beginning in the mid 1980s (*e.g.*, Kilss and Alvey 1985, Carpenter and Fair 1989). A further major stimulus in the U.S. has been the effort to study undercoverage in the 1990 Decennial Census (*e.g.*, Winkler and Thibaudeau 1991). The seminal book by Newcombe (1988) has also had an important role in this ferment.

3. SIMULATION SETTING

The intent of our simulations is to use matching scenarios that are worse than what some users will encounter and to use quantitative data that is both easy to understand and difficult to use in matching.

3.1 Matching Scenarios

For our simulations, we considered four matching scenarios, as in our earlier work (Scheuren and Winkler 1993). The basic idea was to generate data having known distributional properties, adjoin the data to two files that would be matched, and then to evaluate the effect of increasing amounts of matching error on analyses. We started with two files (of size 12,000 and 15,000) having good matching information and for which true match status was known. About 10,000 of these were true matches (before introducing errors) -- for a rate on the smaller or base file of about 83%.

We then generated quantitative data with known distributional properties and adjoined the data to the files. As we conducted the simulations, a range of error was introduced into the matching variables, different amounts of data were used for matching, and greater deviations from optimal matching probabilities were allowed. These variations are described below and shown in figure 1. For each scenario in the figure, the match weight, the logarithm of **R**, is plotted on the horizontal axis with the frequency, also expressed in logs, plotted on the vertical axis. Matches (or true links) appear as asterisks (*), while nonmatches (or true nonlinks) appear as small circles (o):

> Good Scenario (figure 1a). -- Previously, we had concluded that no adjustments for matching error are necessary here. This scenario can happen in systems designed for matching, having good matching variables, and that use advanced matching algorithms. Systems with Social Security Numbers (SSN's) or Employer Identification Numbers (EIN's) might be real world examples. In any event, the true mismatch rate here was under 2%.
>
> Mediocre Scenario (figure 1b). --The mediocre matching scenario consisted of using last name, first name, middle initial, two address variations, apartment or unit identifier, and age. Minor typographical errors were introduced independently into one seventh of the last names and one fifth of the first names. Matching probabilities were chosen to deviate from optimal but considered consistent with those that might be selected by an experienced computer matching expert. An example of a possible real world parallel might be the matching of a file of out of date business information to a computerized file taken from the Yellow Pages. The true mismatch rate here was 6.8%.
>
> First Poor Scenario (figure 1c). -- The first poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names. Moderately severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was for them to be selected in a manner that a practitioner might choose after gaining only a little experience. The true mismatch rate here was 10.1%.
>
> Second poor Scenario (figure 1d). --The second poor matching scenario consisted

of using last name, first name, and one address variation. Minor typographical errors were introduced independently into one third of the last names and one third of the first names. Severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was to represent situations that often occur with lists of businesses in which the linker has little control over the quality of the lists. The true mismatch rate was 14.6%.

With the various scenarios, our ability to distinguish between true links and true nonlinks differs significantly. For the good scenario, we see that the scatter for true links and nonlinks is almost completely separated (Figure 1a). With the mediocre scheme, the corresponding sets of points overlap moderately (Figure 1b); with the first poor scenario, the overlap is substantial (Figure 1c); and, with the second poor scheme, the overlap is almost total (Figure 1d).

RL true mismatch error rates can be reasonably well estimated by the procedure of Belin and Rubin (1995), except in the second poor scenario where the Belin-Rubin procedure would not converge. In practice, for this scenario there is almost no part of the data for which true link status would be known without followup operations. Until now an analysis based on the second poor scenario would not have seemed even remotely sensible. As we will see in Section 4, something of value can be done, even in this case.

3.2 Quantitative Scenario

Having specified the above linkage situations, we then used SAS to generate ordinary least squares data under the model $\mathbf{Y} = \mathbf{6}\ \mathbf{X} + \epsilon$. The **X** values were chosen to be uniformly distributed between 1 and 101 and the error terms ϵ are normal and homoscedastic with variance 35000 -- all such that the regression of **Y** on **X** has an R^2 value in the true matched population of 43%. Matching with quantitative data is difficult because, for each record in one file, there are hundreds of records having quantitative values that are close to the record that is a true match. Additionally, to make modelling and analysis much more difficult, we used all false matches and only 5% of the true matches. We only present the results for the second poor scenario because these are far and away the most dramatic.

See figure 2a for the actual true regression relationship and related scatterplot, as they would appear if there were no matching errors. Note all of the mismatches are plotted but only 5% of the true matches are being used. This has been done to keep the true matches from dominating the results so much that no movement can be seen. Second, in this figure and throughout the remaining ones, the true regression line is always given for reference. Finally, the true population slope or **beta** coefficient (at 5.85) and the R^2value (at 43%) are provided for the data being displayed.

4. SIMULATION RESULTS

We present graphs and results of the recursive process for the second poor scenario. The regression results for two cycles are given in the first two subsections. In the third section, we present results that help explain why such a dramatic improvement can occur.

4.1 First Cycle Results

4.1.1 Regression after Initial **RL $\Rightarrow$RA** Step. -- In figure 2b, we are looking at the regression on the actual observed links -- not what should have happened in a perfect world but what did happen in a very imperfect one. Unsurprisingly, we see only a weak regression relationship between **Y** and **X**. The observed slope or **beta** coefficient differs greatly from its true value (2.47 v. 5.85). The fit measure is similarly affected, falling to 7% from 43%.

4.1.2 Regression after Combined **RL$\Rightarrow$RA$\Rightarrow$EI $\Rightarrow$RA** Step. -- Figure 2c completes our display of the first cycle of our recursive process. Here we have edited the data in the plot displayed as follows. First, using just the 99 cases with a match weight of 3.00+, an attempt was made to improve the poor results given in figure 2b. Using this provisional fit, predicted values were obtained for all the matched cases; then outliers with residuals of 280 or more were removed and the regression refit on the remaining pairs. This new equation was essentially $\mathbf{Y} = 4.5\mathbf{X} + \epsilon$ with a variance of 40000. Using our earlier approach (Scheuren and Winkler 1993), a further adjustment was made in the **beta** coefficient from 4.5 to 5.4. If a pair of matched records yielded an outlier, then predicted values using the equation $\mathbf{Y} = 5.4\mathbf{X}$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

4.2 Second Cycle Results

4.2.1 True regression (for reference). -- Figure 3a displays a scatterplot of **X** and **Y**, as they would appear if they could be true matches based on a second **RL** step. The second **RL** step employed the predicted **Y** values as determined above; hence it had more information on which to base a linkage. This meant that a different group of linked records was available after the second **RL** step. In particular, since a considerably better link was obtained, there were fewer false matches; hence our sample of all false matches and 5% of the true matches dropped from 1104 in figures 2a thru 2c to 656 for figures 3a thru 3c. In this second iteration, the true slope or **beta** coefficient and the R^2 values remained, though, virtually identical for the slope (5.85 v. 5.96) and fit (44% v. 45%).

4.2.2 Regression after second **RL $\Rightarrow$RA** Step. -- In figure 3b, we see a considerable improvement in the relationship between **Y** and **X** using the actual observed links after the second **RL** step. The slope has risen from 2.47 initially to 5.07 here. Still too small but much improved. The fit has been similarly affected, rising from 7% to 35%.

4.2.3 Regression after Combined **RL$\Rightarrow$RA$\Rightarrow$EI $\Rightarrow$RA** Step. -- Figure 3c completes the display of the second cycle of our recursive process. Here we have edited the data as follows. First, using just the 54 cases with a match weight of 7.00+, an attempt was made to further improve on the results obtained in figure 3b.

Using this fit, another set of predicted values was obtained for all the matched cases. This new equation was essentially $\mathbf{Y} = 5.5\mathbf{X} + \epsilon$ with a variance of about 35000. Using our earlier approach (Scheuren and Winkler 1993), a further adjustment was made in the **beta** coefficient from 5.5 to 6.0. Again, if a pair of matched records yields an outlier, then predicted values using the equation **Y** = 6.0**X** were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value. The plot in figure 3c gives the adjusted values.

4.3. Further Results

While figures 2a-2c and 3a-3c summarize the dramatic improvement in the fitted model after the second matching pass, we need additional results that will help us understand why the improvement occurred. Table 1 summarizes the numbers of true and false matches on the two matching passes. It provides the template for the other results of this subsection. Comparing the numbers from the first two passes, we see that when approximately 8800 true matches are accumulated, then the first pass has 356 false matches while the second pass has only 152. Figure 4 shows the how the second pass improves over the second pass (figure 1d).

Table 1. True and False Matches by Weight by Matching Pass

	First Pass			
	Number		Cumulative Number	
Weight	Trues	Falses	Trues	Falses
3+	1223	28	1223	28
2	2997	72	4220	100
1	3276	85	7496	185
0	1330	171	8826	356
-1	365	289	9191	645
-2	63	388	9254	1033
-3	2	226	9256	1259
	Second Pass			
	Number		Cumulative Number	
Weight	Trues	Falses	Trues	Falses
7	946	13	946	13
6	1697	23	2643	36
5	2166	38	4809	74
4	1361	22	6170	96
3	1260	14	7430	110
2	897	24	8327	134
1	505	18	8832	152
0	293	78	9125	230
-1	217	136	9342	366
-2	113	129	9455	495
-3	34	67	9489	562

Detailed review of the true and false matches during the first and second passes shows that approximately 150 pairs that are false on the first pass are true on the second while 50 pairs that are true on the first pass are false on the second. Many of the 50 pairs that are false on the second pass have true quantitative values that are outliers. Approximately 100 pairs among the first 366 false matches are associated records that can never be correctly matched because the true corresponding match is not present.

Figures 5a and 5b are plots of the true matches (5% sample) and false matches (100% of cases) in the last scenario that we used for modelling and analysis. The graphs illustrate why we need the adjustment procedures of our earlier paper (Scheuren and Winkler 1993). The true matches have a **beta** coefficient 6.26 with R^2 of 0.50 while the false matches have a **beta** coefficient 4.02 with R^2 of 0.32. Those false matches that have quantitative values relatively close to the modelled values and that are not picked up by our outlier detection method have a tendency to distort the distribution.

5. CONCLUSIONS AND AREAS FOR FUTURE STUDY

In principle, the recursive process of matching and modelling could have continued. Indeed, while we did not show it in this paper, the **beta** coefficient of our example did not change much during a third matching pass.

At first it would seem that we should be happy with the results. They take a seemingly hopeless situation and give us a fairly sensible answer. A closer examination, though, shows a number of places where the approach taken is weaker than it needs to be or simply unfinished.

5.1 False Matches and Use of Belin-Rubin Procedures

We are most attracted to an idea of Howard Newcombe's involving the use of a sample of **known** nonmatches. In our earlier work (Scheuren and Winkler 1993) on this problem, the matching we proposed involved getting two links for each case in the base file. The second link would be a next best and usually could be assumed to be a false match. How could these second (false) matches be used to help to improve the modelling?

If the matching is good enough so that the Belin-Rubin algorithms work (Belin and Rubin 1995), we can calculate a true link probability for each match. We would then be able to estimate the number of false links among our best matched cases. This number of cases could be selected from the second best match file -- perhaps simply at random or better in a balanced way (such that, say, the means of **X** and **Y** in this false matched sample file agreed with the corresponding values in the original or best match file). A possible next step here would be to match the false matched sample to the original best matched cases and remove the "closest" pairs. This would be done instead of looking for outliers and removing all those at some distance from the regression fit of the data (as was described in 4.1.2).

Even if the Belin-Rubin algorithms do not converge on the first cycle, Newcombe's idea of using a file of nonmatches might still be tried once the recursive process yielded matches of sufficient quality to employ it. In the present example this would have been possible at the second cycle, even though it was not possible initially.

There is still another way to use Newcombe's idea that we would like to try. Rather than obtaining the joint **X,Y** distribution directly using the recursive algorithm $\mathbf{RL} \Rightarrow \mathbf{RA} \Rightarrow \mathbf{EI}$

⇒**RA,** we might focus on the distribution and only indirectly on the data. For example, after the first cycle through **RL⇒RA⇒EI ⇒RA**, the second cycle might be **RL⇒RA⇒EM ⇒RA** where an **EM** algorithm replaces the Edit/Imputation **EI** step. Here the **EM** step is introduced to implicitly deal with the mixture of true and false matches using the proportion of false matches from Belin-Rubin as the mixing constant. In any event, however we approach the estimation, the combination of Newcombe's ideas and the use of EM-based parameter estimation techniques to record linkage seems very promising. (e.g., Winkler 1994, Meng and Rubin 1994).

5.2 Generalizability Concerns

We have looked at a simple regression of one variable from one file with another variable from another. What happens when this is generalized to the multiple regression case? We are working on this now and sensible results are starting to emerge which have given us insight into where further research is required. There is also the case of multivariate regression. Here the problem is harder and will be more of a challenge:

> First, to make use of multivariate data, we need to have better ways of modelling it than the simple method of this paper. The likely best methods will be variants and extensions of Little and Rubin (1987, Chapters 6 and 8) in which predicted multivariate data has important correlations accounted for. If we take two variables from one file and two from another, then can we make use of the fact the two variables taken from one file have the correct two-variable distribution but may be falsely matched. To handle this, new software for Little-Rubin methods and a multivariate generalization of Scheuren-Winkler (1993) need to be written.
>
> Second, we have not yet developed effective ways of utilizing the predicted and unpredicted quantitative data. Simple multivariate extensions of the univariate comparison of **Y** values in this paper do not seem to work. The additional distinguishing power of the multivariate quantitative data in comparison with name and address information needs to be accounted for. The EMH methods of Winkler (1994) and the MCECM methods of Meng and Rubin (1994) may be useful in this regard. The existing matching software needs enhancement too. EMH software (Winkler 1994) is currently available but the precise methods of best applying remain to be determined.

On some other issues we plan to conduct more simulations to inform our intuitions. For example, what happens when the relationship between **Y** and **X** is weak in the population. Maybe, then, we cannot improve the match enough to make all the work being done here worthwhile? Our suspicion here is that there is some threshold on R^2 below which the gains in matching strength do not compensate enough for poor matching variables to make the methods we are advocating serviceable. So far we have only looked seriously at two scenarios -- $R^2 = .78$ (Winkler and Scheuren 1995) and $R^2 = .43$ (in the present paper). What happens if R^2 drops still further to, say, $R^2 = .22$? Our efforts still might be worthwhile but as a way perhaps to ascertain rough bounds on the **beta** coefficient rather than to obtain a point estimate that might have a conventional interpretation.

Many other questions remain too. In particular, what happens when the overlap between the two files is very low (it was high in our example)? Here we have yet to attempt a simulation but will soon. Looking directly at the estimated standard errors of the beta coefficient is planned too, as a function, say, of the root mean square error.

5.3 Statistical Technology and Statistical Theory

While the mathematical foundations of **RL** (Fellegi and Sunter 1969) and **EI** (Fellegi and Holt 1976) appeared quite a while ago, development of these powerful statistical tools was hampered because **RL** is primarily a computer science problem and **EI** is primarily an operations research problem. The **RL** and **EI** technologies are now mature enough to begin to fully exploit as statistical inference tools. It is our current view that no additional advances in computer and operations research are now needed for the methods that we described here to be generally applied.

This is not to say that no more statistical theory is needed for dealing with the problems addressed here. In fact, the paper has been mainly about technological possibilities. Our discussion has not been independent of theoretical considerations; but, conversely, the theoretical underpinnings of the ideas being explored have not all been worked out either. This early, intuitive approach is not unexpected of work in progress. We do not apologize for it; rather, in some ways it allows you, the listener (or reader), to become a player. One of our goals in our earlier work was to get others involved. It continues to be our goal -- whether that involvement be on the theoretical side or through an application.

REFERENCES

BELIN, T. R., and RUBIN, D. B. (1995), "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association*, **90**, 694-707.

CHERNIKOVA, N.V. (1964) "Algorithm for Finding a General Formula for the Nonnegative Solutions of a System of Linear Equations," *USSR Computational Mathematics and Mathematical Physics*, 4, 151-158.

CHERNIKOVA, N.V. (1965) "Algorithm for Finding a General Formula for the Nonnegative Soilutions of a System of Linear Equations," *USSR Computational Mathematics and Mathematical Physics*, 5, 228-233.

CARPENTER, and FAIR, M.. (Editors) (1989), *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Epidemiological Research Conference, in Ottawa, Ontario, Canada, August 30-31, 1989, Statistics Canada.

FELLEGI, I. and HOLT, T.(1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the of the American Statistical Association*, **71**, 17-35.

FELLEGI, I., and SUNTER, A. (1969), "A Theory of Record Linkage," *Journal of the of the American Statistical Association*, **64**, 1183-121

FILION, J. and SCHIOPU-KRATINA, I. (1993), "On the Use of Chernikova's Algorithm for Error Localization," *Statistics Canada Technical Report*.

GRANQUIST, L. (1984), "On the Role of Editing," *Statistic Tidshrift*, **2**, 105-118.

JABINE, T. B. and SCHEUREN, F. J. (1986), "Record linkages for statistical purposes: methodological issues," *Journal of Official Statistics*, **2**, 255-277.

JARO, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414-420.

KILSS, B. and ALVEY, W. (Editors) (1985), *Record Linkage Techniques- 1985*, U.S. Internal Revenue Service, Publication 1299 (2-86).

KIM, J. J., and WINKLER, W. E. (1995), "Masking Microdata Files," *American Statistical Association, Proceedings of the Section of Survey Research Methods*, to appear.

KOVAR, J. G., WHITRIDGE, P., and MACMILLAN, J. (1988), "Generalized Edit and Imputation System for Economic Surveys at Statistics Canada," *American Statistical Association, Proceedings of the Section of Survey Research Methods*, 627-630.

LITTLE, R. J. A. and RUBIN, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

MENG, X., and RUBIN, D. B. (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, **80**, 267-278.

NEWCOMBE, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

NEWCOMBE, H. B., KENNEDY, J. M., AXFORD, S. J., and JAMES, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.

NEWCOMBE, H., FAIR, M., and LALONDE, P., (1992), "The Use of Names for Linking Personal Records," *Journal of the American Statistical Association*, **87**, 1193-1208.

NETER, J., MAYNES, E. S., and RAMANATHAN, R. (1965), "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, **60**, 1005-1027.

NORDBOTTEN, S. (1963), " Automatic Editing of Individual Observations," presented at the Conference of European Statisticians, U.N. Statistical and Economic Commission of Europe.

OH. H. L. and SCHEUREN, F. (1975) "Fiddling Around with Mismatches and Nonmatches," *American Statistical Association Proceedings, Social Statistics Section.*

RUBIN, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

RUBIN, D. S. (1975), Vertex Generation and Cardinality Constrained Linear Programs," *Operations Research*, **23**, 555-565.

SANDE, I. (1982), "Imputation in Surveys: Coping with Reality, *The American Statistician*, 145-182.

SCHEUREN, F., and WINKLER, W. E. (1993), "Regression Analysis of Data Files that are Computer Matched," *Survey Methodology*, **19**, 39-58.

TEPPING, B. (1968), "A Model for Optimum Linkage of Records," *Journal of the American Statistical Association*, **63**, 1321-1332.

WINKLER, W. E. (1994), "Advanced Methods of Record Linkage," *American Statistical Association, Proceedings of the Section of Survey Research Methods*, 467-472.

WINKLER, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al*. (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.

WINKLER, W. E., and DRAPER, L. (1996), "Application of the SPEER Edit System," in *Statistical Data Editing, Volume 2,* U.N. Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, to appear.

WINKLER, W. E., and PETKUNAS, T. (1996), "The DISCRETE Edit System," in *Data Editing, Volume 2*," U.N. Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, to appear.

WINKLER, W. E. and SCHEUREN, F. (1995), "Linking Data to Create Information," *Proceedings of Statistics Canada Symposium 95*.

WINKLER, W. E. and THIBAUDEAU, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," Statistical Research Division Technical Report, U.S. Bureau of the Census.

Figure 1a. Good Matching Scenario

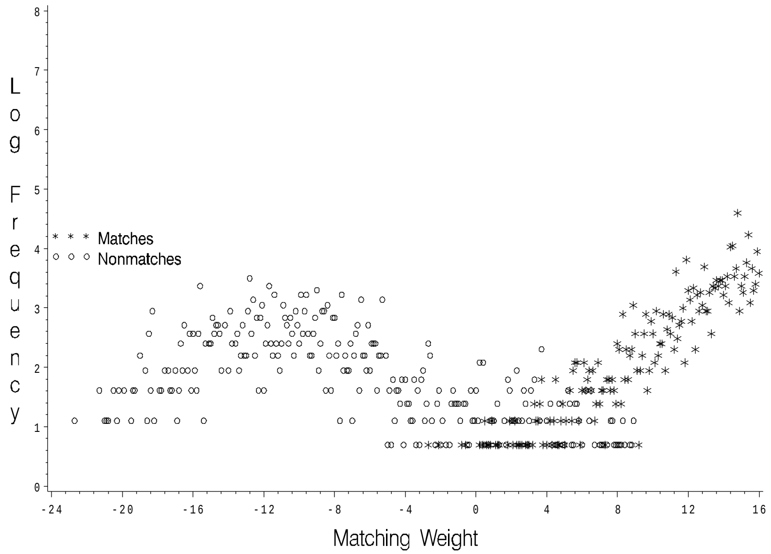


Figure 1b. Mediocre Matching Scenario

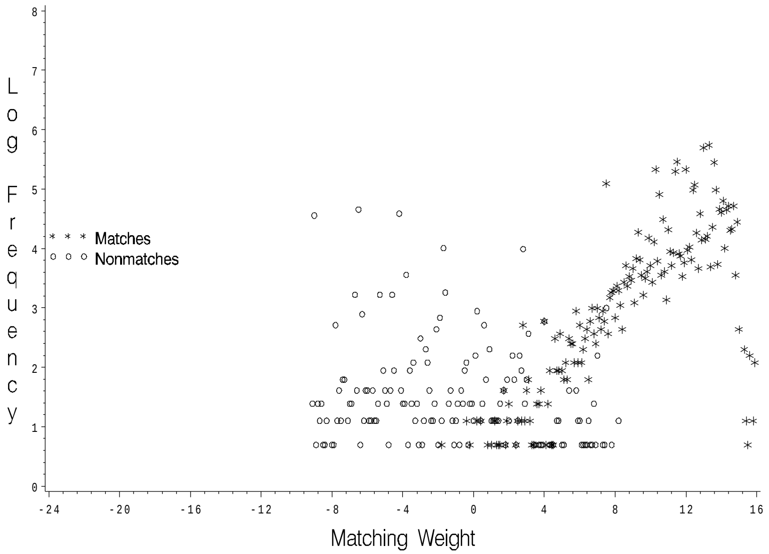


Figure 1c. 1st Poor Matching Scenario

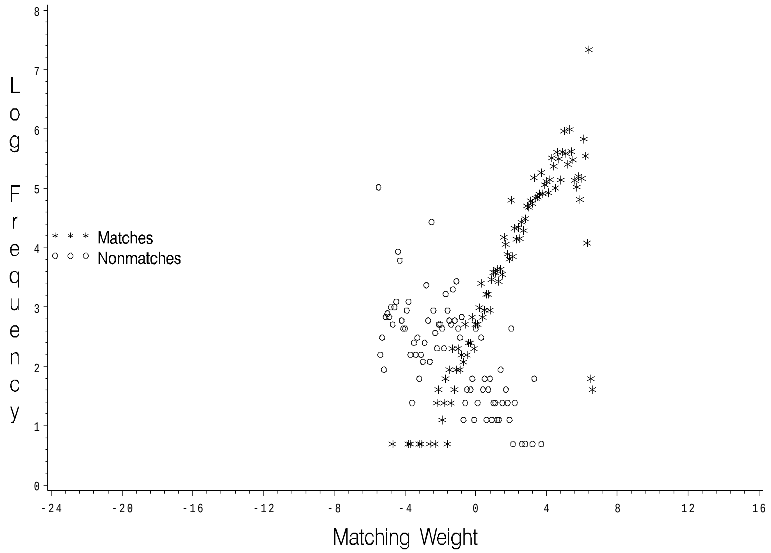


Figure 1d. 2nd Poor Matching Scenario

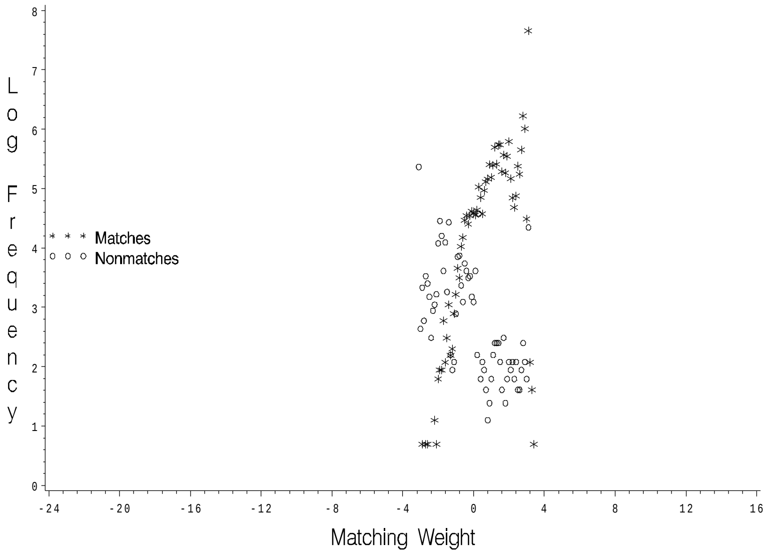


Figure 2a. 2nd Poor Scenario, 1st Pass

All False & 5% True Matches, True Regression Data

1104 Points, beta=5.85, R−square=0.43

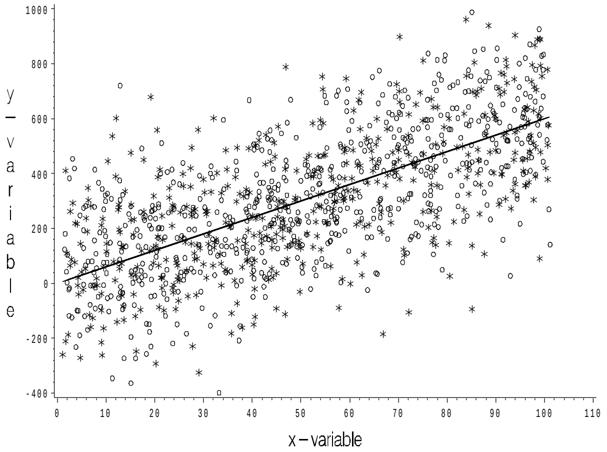


Figure 2b. 2nd Poor Scenario, 1st Pass

All False & 5% True Matches, Observed Data

1104 Points, beta=2.47, R−square=0.07

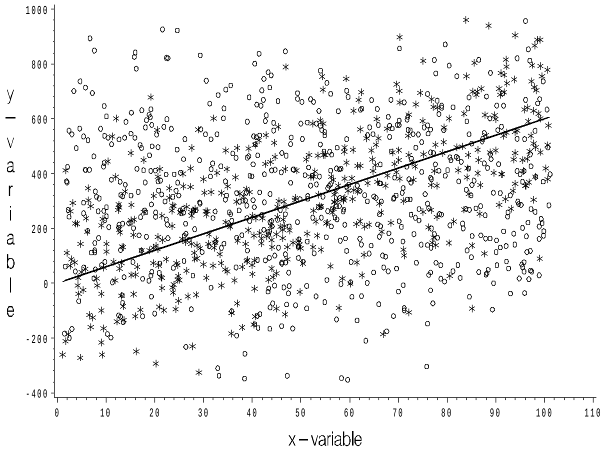


Figure 2c. 2nd Poor Scenario, 1st Pass

All False & 5% True Matches, Outlier−Adjusted Data

1104 Points, beta=5.21, R−square=0.47

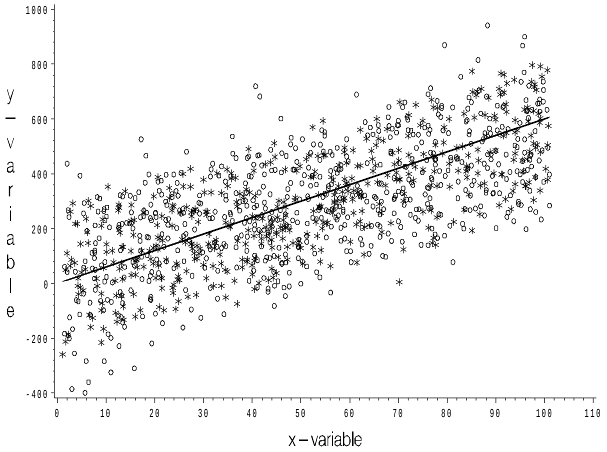


Figure 3a. 2nd Poor Scenario, 2nd Pass

All False & 5% True Matches, True Regression Data

656 Points, beta=5.96, R−square=0.45

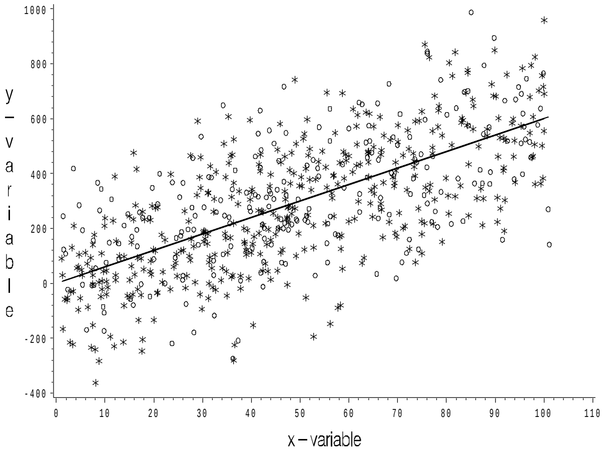


Figure 3b. 2nd Poor Scenario, 2nd Pass

All False & 5% True Matches, Observed Data

656 Points, beta=5.07, R−square=0.35

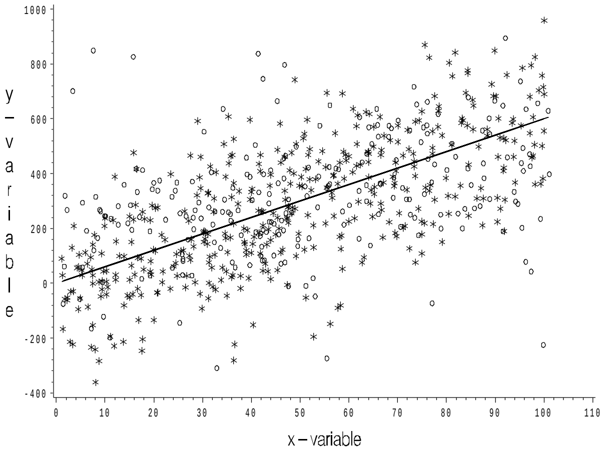


Figure 3c. 2nd Poor Scenario, 2nd Pass

All False & 5% True Matches, Outlier−Adjusted Data

656 Points, beta=5.50, R−square=0.44

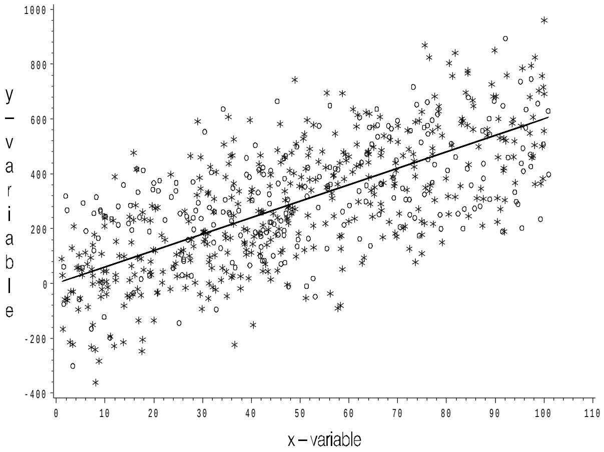


Figure 4. Log of Frequency vs Weight, 2nd Pass

2nd Poor Matching Scenario, Matches=* vs Nonmatches=o

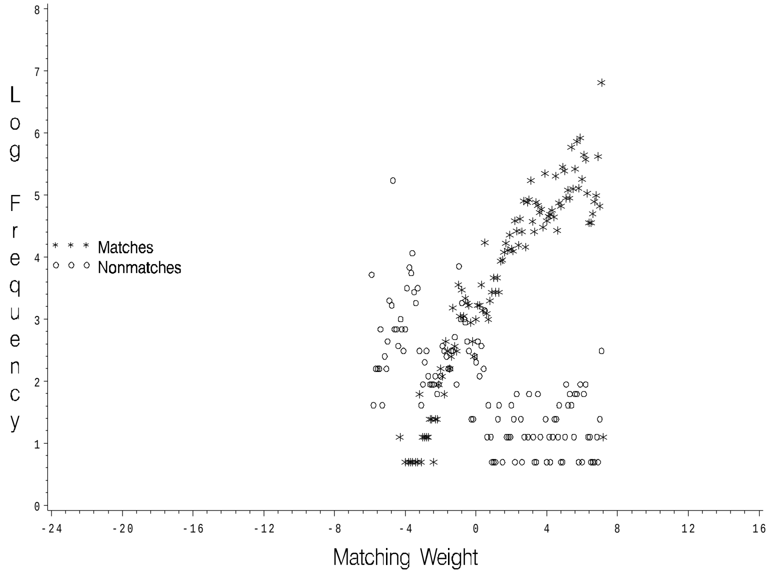


Figure 5a. 2nd Poor Scenario, 2nd Pass

True Matches (5%), Outlier-Adjusted Data

426 Points, beta=6.26, R-square=0.50

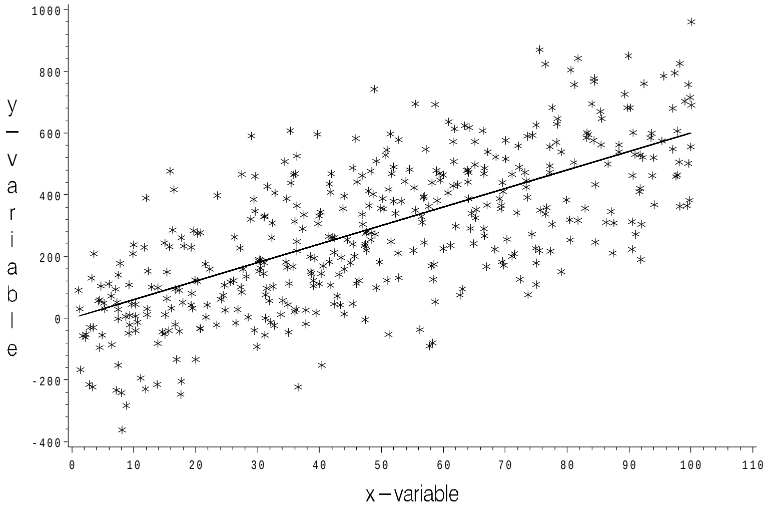
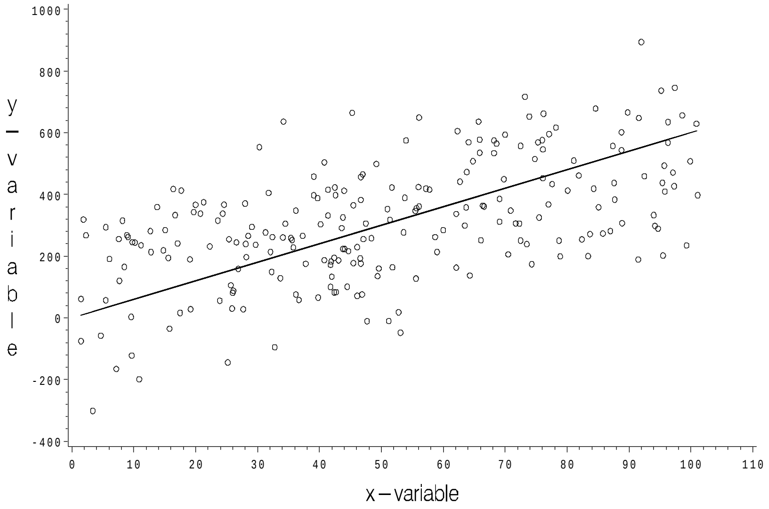


Figure 5b. 2nd Poor Scenario, 2nd Pass

False Matches, Outlier-Adjusted Data

230 Points, beta=4.02, R-square=0.32


y-variable
x-variable